Fibre Channel and IP SAN Integration

Dr. Joseph L White

(*was Henry Yang)*

(My presentation is based on his conference paper)

McDATA Corporation

4 McData Parkway, Broomfield CO 80021-5060

[3850 N. First Street, San Jose, CA 95134]

+1-408-519-3744

Joe.White@McDATA.com

NASA/IEEE MSST 2004

12th NASA Goddard/21st IEEE Conference on
Mass Storage Systems & Technologies

The Inn and Conference Center
University of Maryland University College
Adelphi MD USA
April 13-16, 2004

Introduction

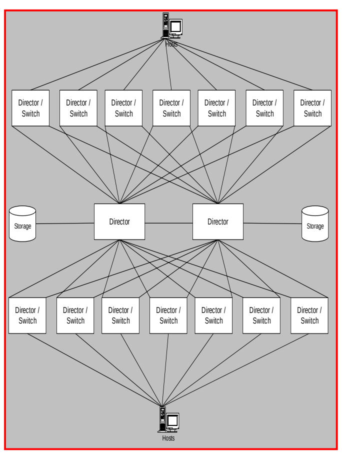
Why look at Converged SANs?

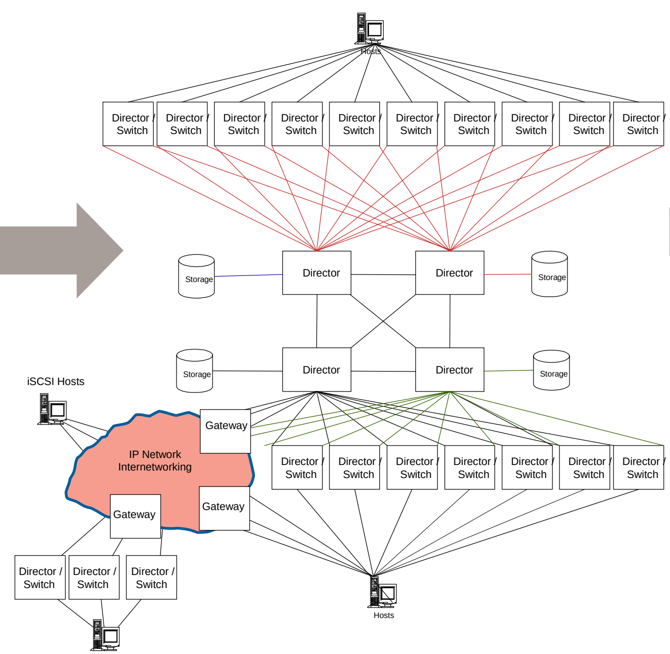
- **IT Professionals and Data Centers are under pressure:**
  – Return on Investment demands
  – Explosive growth
    - of Data and Storage Systems
    - Information Processing Capacity
    - User Demand
  – New Technologies

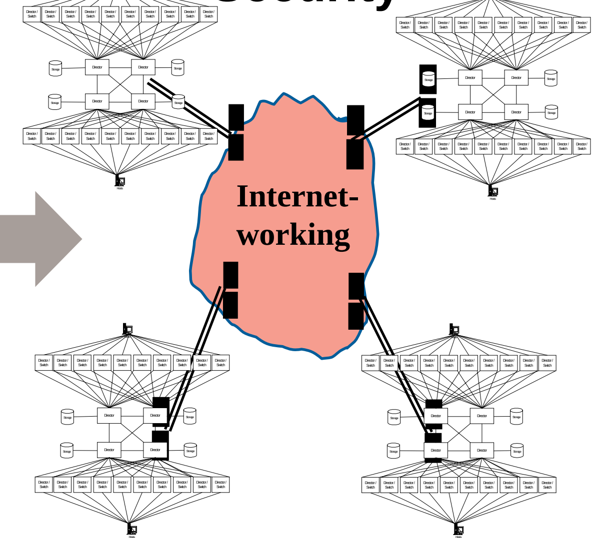
SANs Must Become More Interconnected

The technology exists to take a single FC fabric SAN towards a large scale

Today	Internetworking	Backbone SAN
Large Fabric	Larger Fabrics	Full Connectivity
1000+ ports	5000+ ports	50,000+ ports
FC Only	LAN/MAN/WAN	10G
	Multi-protocols	QoS
		Security







SAN Development

- **To Respond SANs**
  - Must Become
    - More Scalable
    - More Interconnected
  - And Provide
    - More Partitioning
    - More Provisioning
    - More Elegant and Unified Management

SANs Requirements

- **Technology pushed by mission critical applications**
  - Deployments have evolved and matured into multi-Terabit networks
  - Critical to the Data Center
- **Demanding Requirements:**
  - High Throughput
  - Low Latency
  - Wide Scalability
  - Robustness
  - High Availability
- **Questions:**
  - What are the critical system level behaviors in this environment?
  - How do these behaviors interact…
    - As SANs continue to scale?
    - With the introduction of IP Storage interconnects?

IP Storage Protocols

- **Provide Reliable WAN Interconnect for FC**
  - ***FCIP***
    - Fabric Level FC Extension
  - ***iFCP***
    - Device Level FC Extension

- **As an alternative to FC**
  - ***iSCSI***
    - Native IP Storage protocol

- **Centralized discovery and notification service**
  - ***iSNS***
    - Provides a Name Service for IP Storage

- **All Are TCP/IP based**

IP SANs and IP Connected FC SANS

- **Performance Implications of**
  - Long Delays
  - High Bandwidth
  - LFNs
    - Both Long Delay and High Bandwidth
  - Unreliable Networks

- **Protocol Tuning**
  - How can protocols be adjusted to improve performance

- **These areas are critical to FC & IP SAN integration**

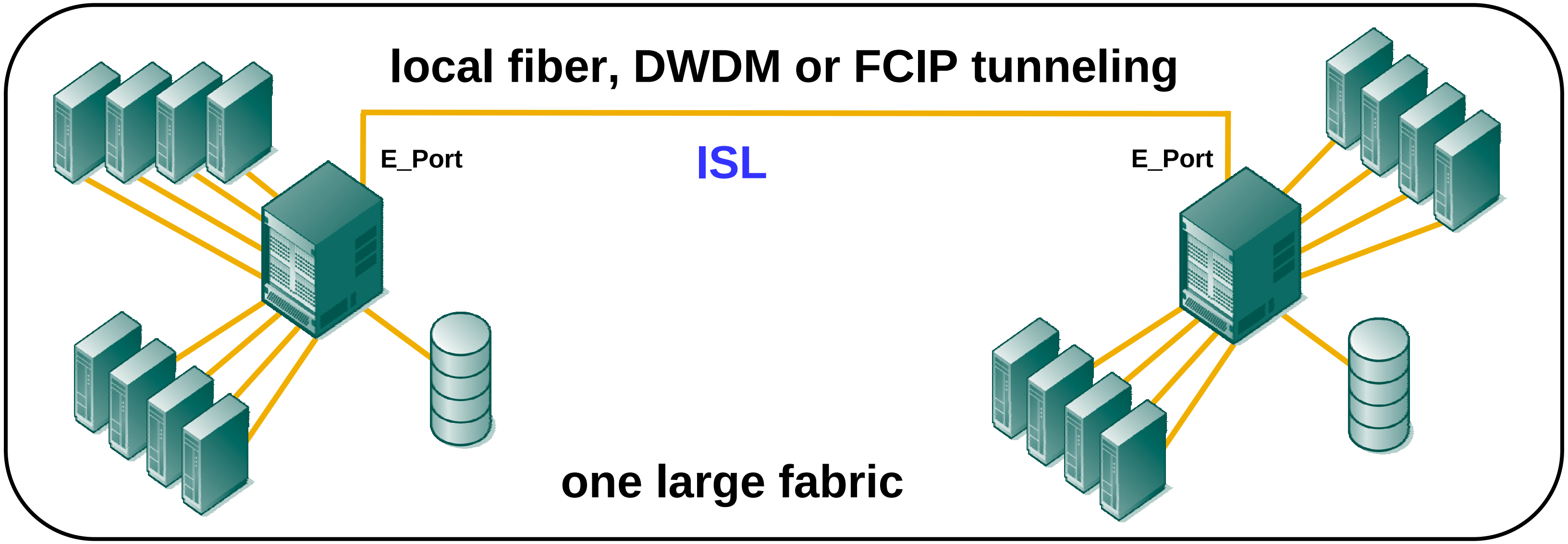
- **Several Case Studies will be covered**
  - Actual Deployments
  - Academic Work

IP & FC Convergence

Deployment Scenarios and Problems Solved Today

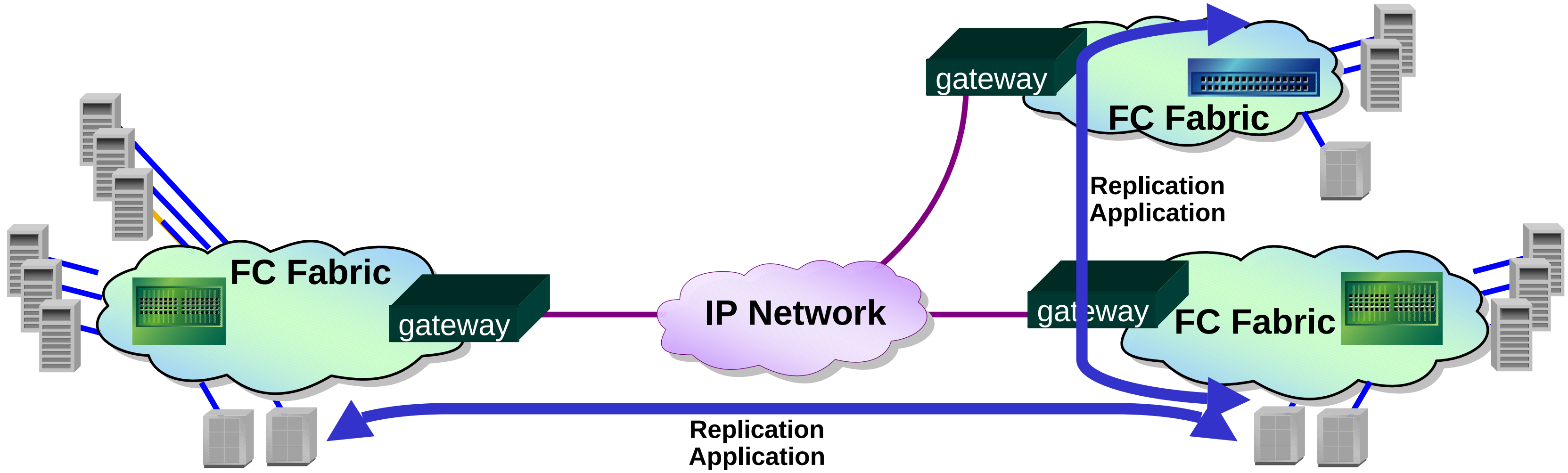
07-APR-2004

SAN Extension



- **Joins Remote Fabrics into one large fabric**
  – Same Benefits and Issues as a single local FC
  – Except for the distance of the extension links

Data Replication, Business Continuance, Branch Consolidation



- **Business Continuance**
  - Cost effective solution for regional offices, remote sites
  - WAN connectivity usually across medium to long latency and slower speed IP networks
  - Normally uses major replication applications
    - EMC SRDF, MirrorView, SAN Copy
    - StorageTek / IBM FAStT RVM
    - XIOtech REDI SANlinks
    - HDS TrueCopy
    - LSI Logic RVM
    - HP DRM

- **Added features can provide higher throughput and lower telecom costs:**
  - Compression, Fast Write, Tape Pipe, etc

Full Internetworking:
Provides Isolation but Allows Connectivity

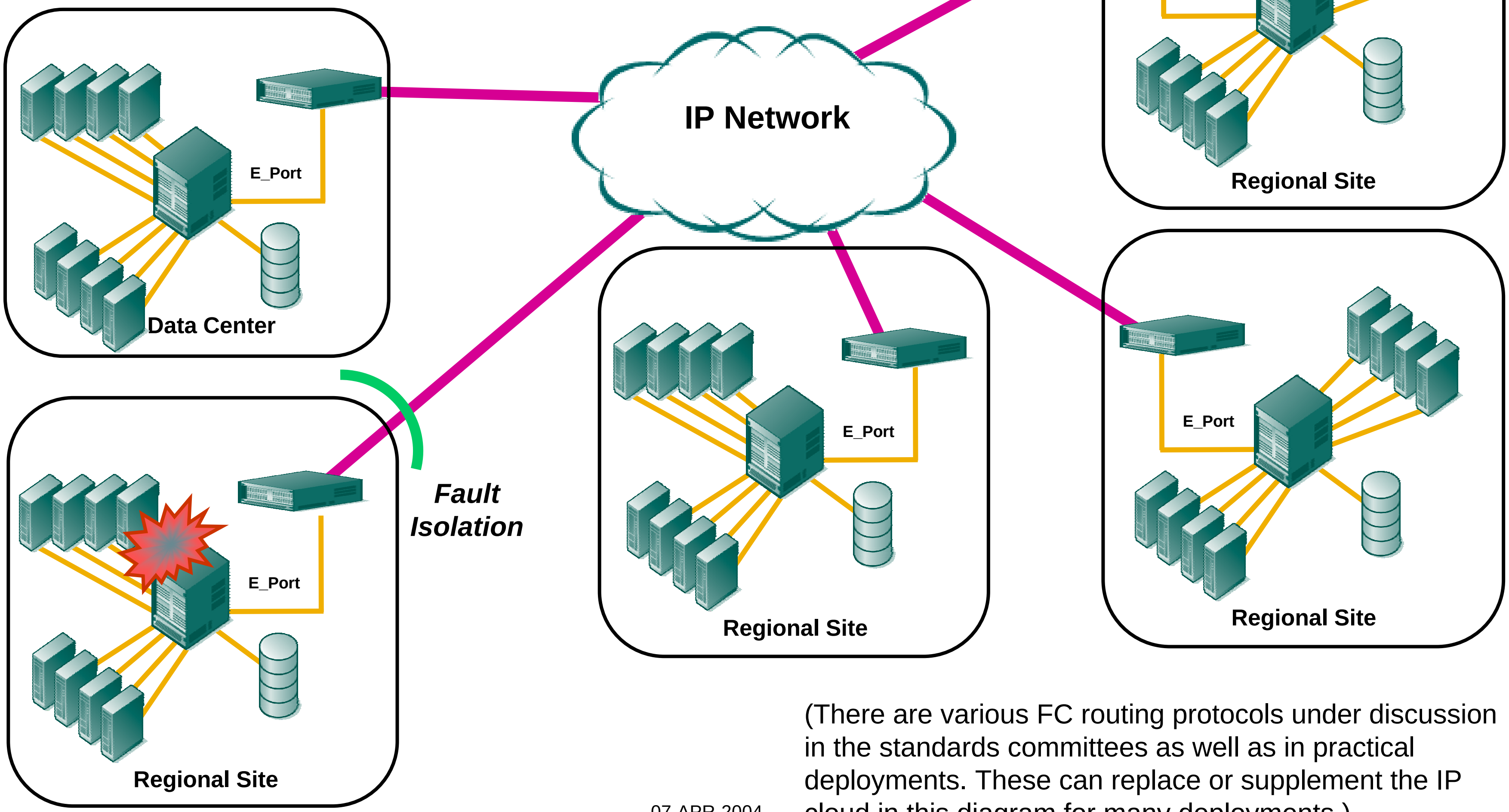
Each box in the diagram is an Independent SAN Fabric with Independent Fabric Services

Reduces Downtime Cost

Provides better asset protection and availability

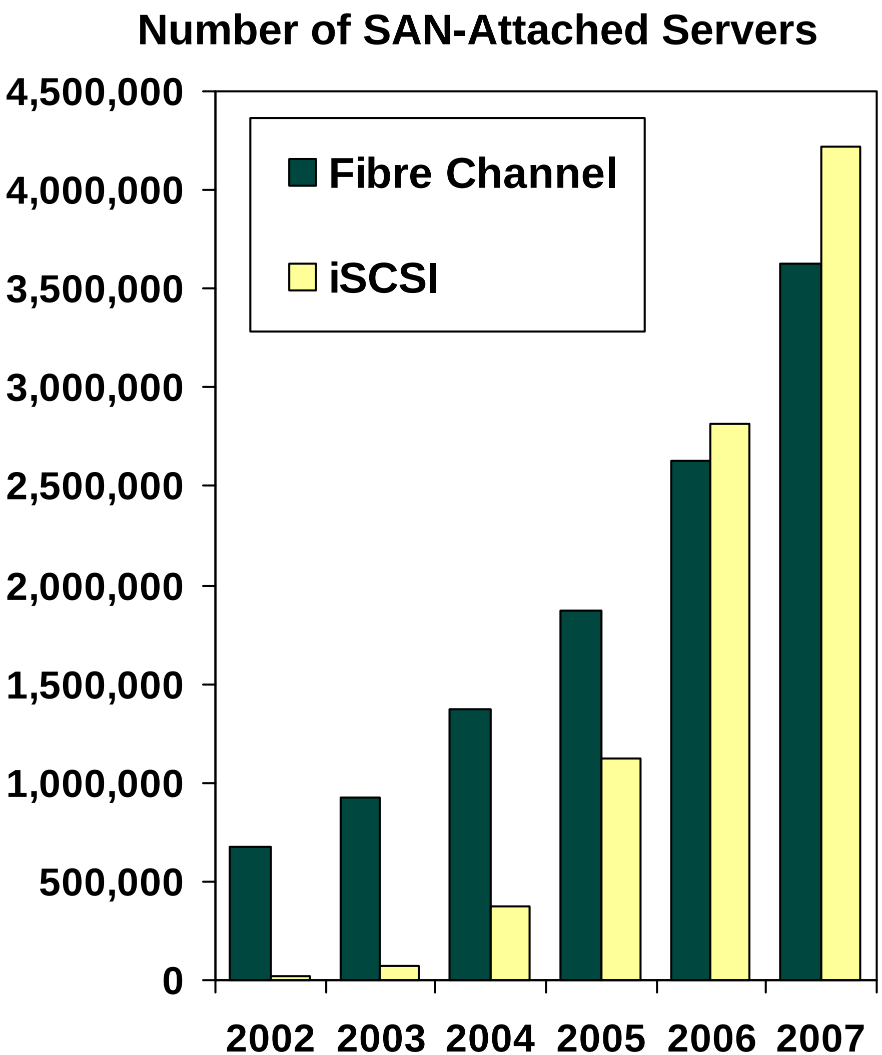
Allows autonomous maintenance regions

Provides security boundaries



(There are various FC routing protocols under discussion in the standards committees as well as in practical deployments. These can replace or supplement the IP cloud in this diagram for many deployments.)

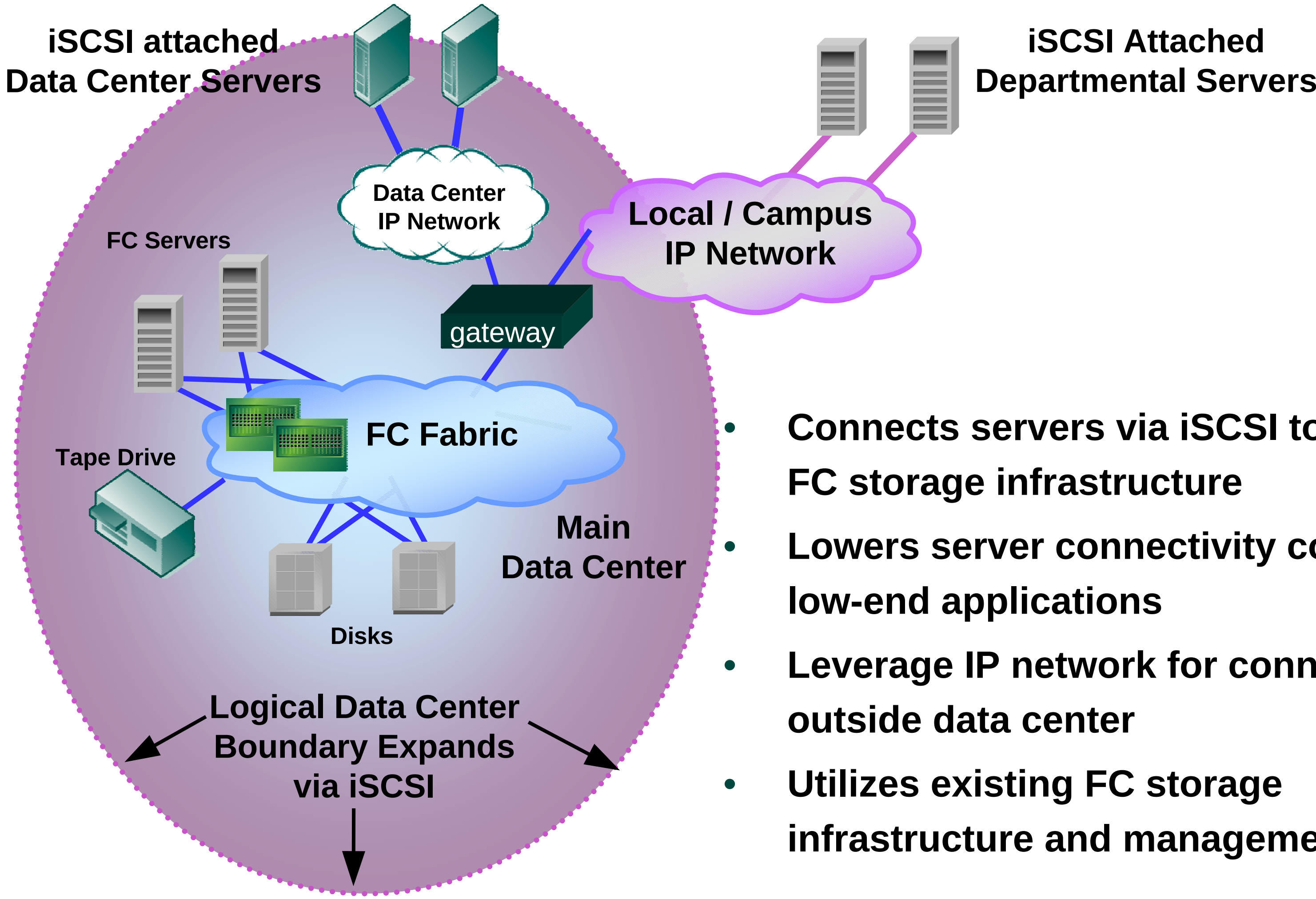
iSCSI Complements FC to Expand SAN Market



Source: Gartner, June 2003

- Majority of iSCSI servers connect to FC storage thru ‘05
- Majority of iSCSI servers use GE NICs and iSCSI OS drivers
- Software iSCSI initiators offer attractive incremental costs for server attachment to SANs
- TOEs and iSCSI HBAs for hardware acceleration enhance performance mainly for native iSCSI SANs

iSCSI: The Distributed Data Center



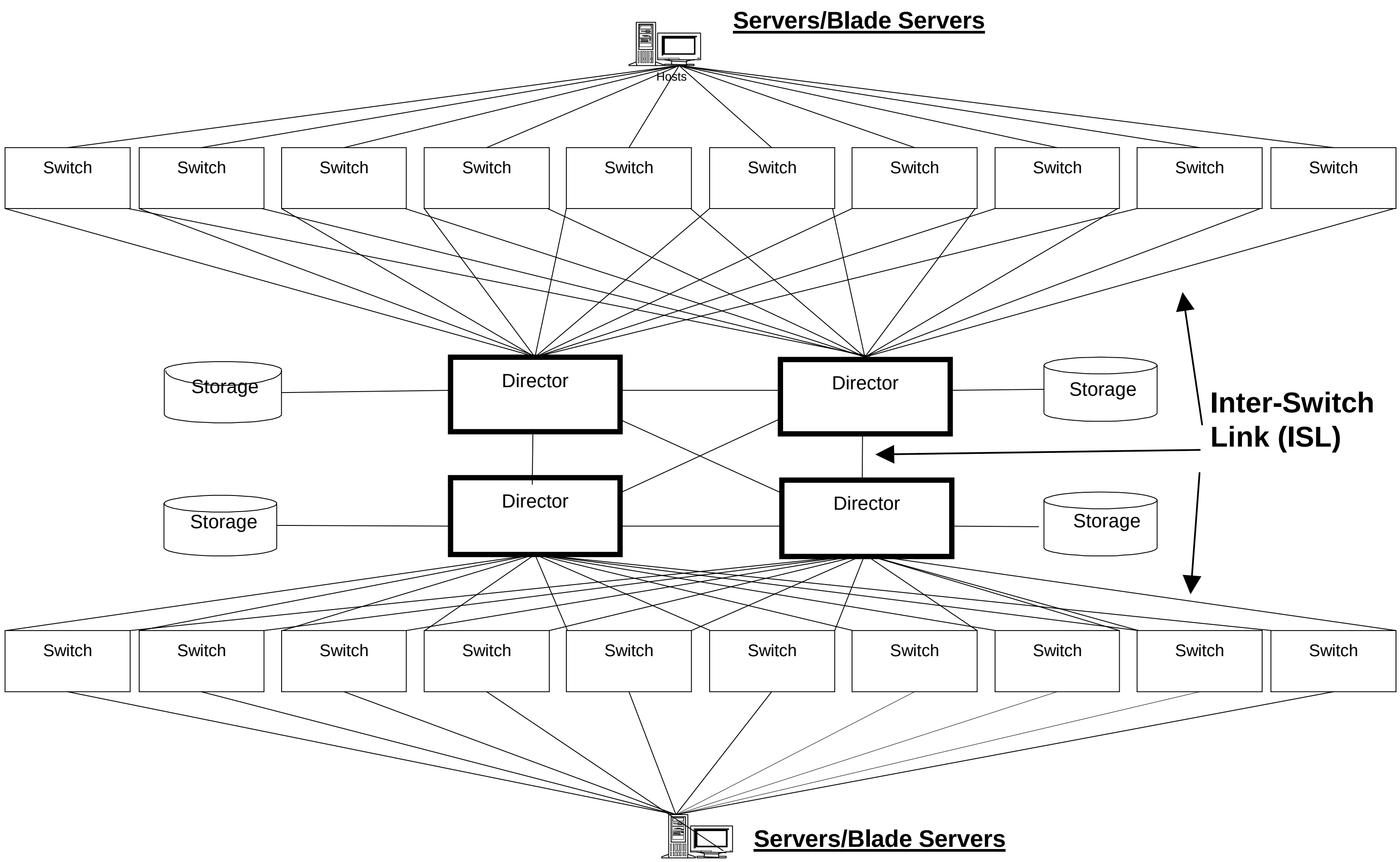
- Connects servers via iSCSI to existing FC storage infrastructure
- Lowers server connectivity cost for low-end applications
- Leverage IP network for connectivity outside data center
- Utilizes existing FC storage infrastructure and management tools

FC SANs

Characteristics and Operation

07-APR-2004

FC SAN Physical Components



FC SAN Physical Components

- **Directors**
  - High Port Density
  - High Aggregate Bandwidth
  - High Availability
  - used in the core of the FC fabric
- **Switches**
  - smaller and less expensive
  - used at edge of FC fabric
- **Devices**
  - Servers & Workstations
  - Storage Devices
    - Discs, Storage Arrays, Tapes
  - Connects to the FC Fabric through a port on a switch or director
  - Allowed connectivity is Any to Any
    - But effective connectivity is Some to Some

FC SAN Physical Components

- **ISLs – Inter-Switch Links**
  - 1G→2G has occurred
  - 4G and 10G on the way
  - Can be extended over distance
    - Campus, MAN and WAN
    - Technologies: T1/T3, ATM, IP, SONET, dark fiber, DWDM
  - Multiple redundant connections used within the fabric for High Availability

- **Each isolated FC Fabric is known as a "SAN Island"**
  - Deployments typically contain more than one SAN Island

FC SAN Logical Components

- **Fabric Initialization**
  - Parameter Exchange
  - Principal Switch Selection
  - Address Assignment
  - Path Computation
  - Zone Merge

- **FSPF (Fabric Shortest Path First)**
  - Used for Path Computation
  - Link state protocol
  - Computes shortest path for frame forwarding

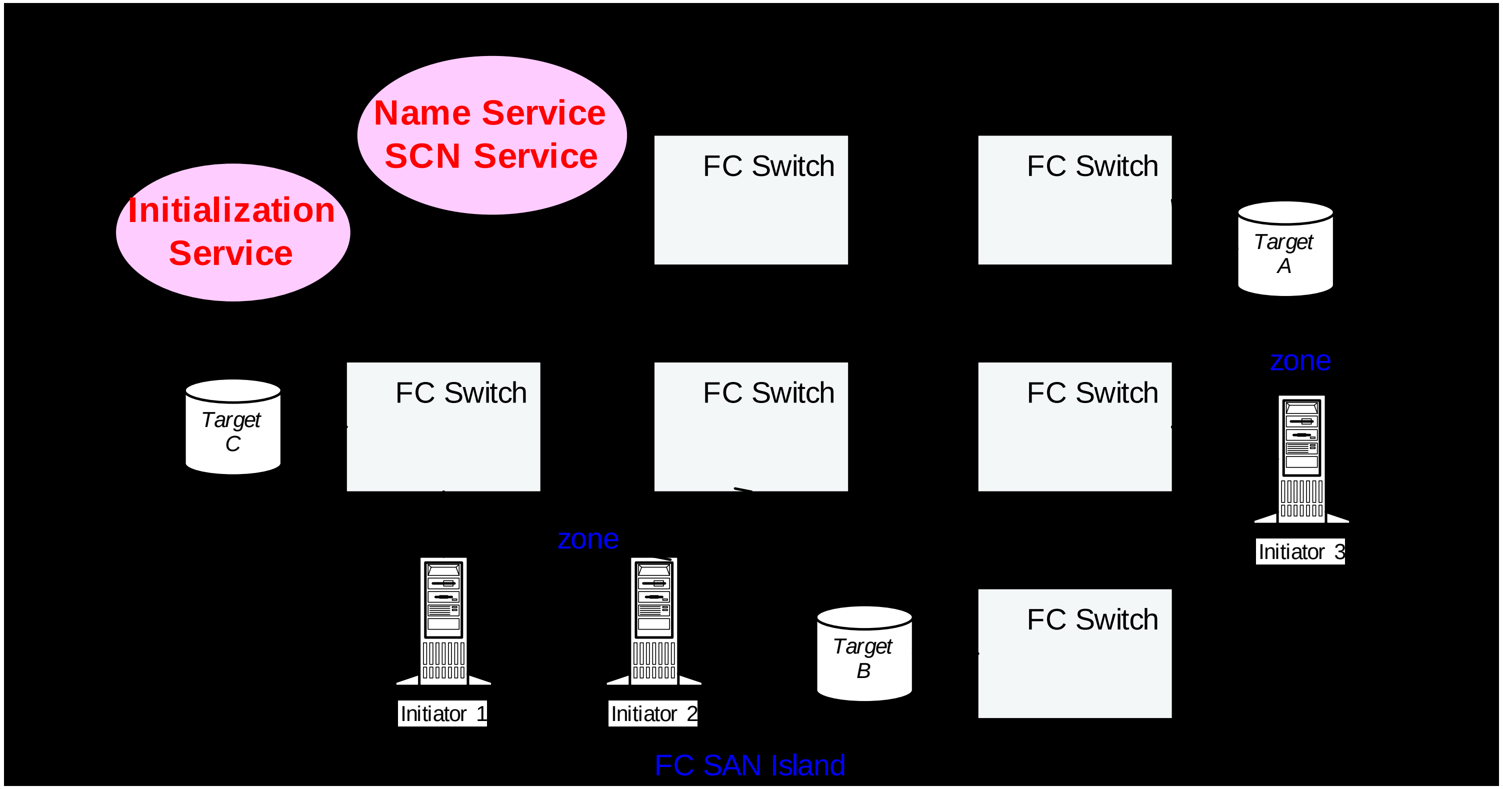
FC SAN Logical Components

- **Name Service & State Change Notification**
  - Resource discovery
  - Configuration
  - Change management

- **FC Zoning**
  - Overlay onto network
  - Limits the discovery time visibility of devices
  - Controls connectivity
  - A single device can be in multiple zones
    - allows for shared access to a single device

FC Zones and Services

Simple FC example in very small fabric



FC SAN Logical Components

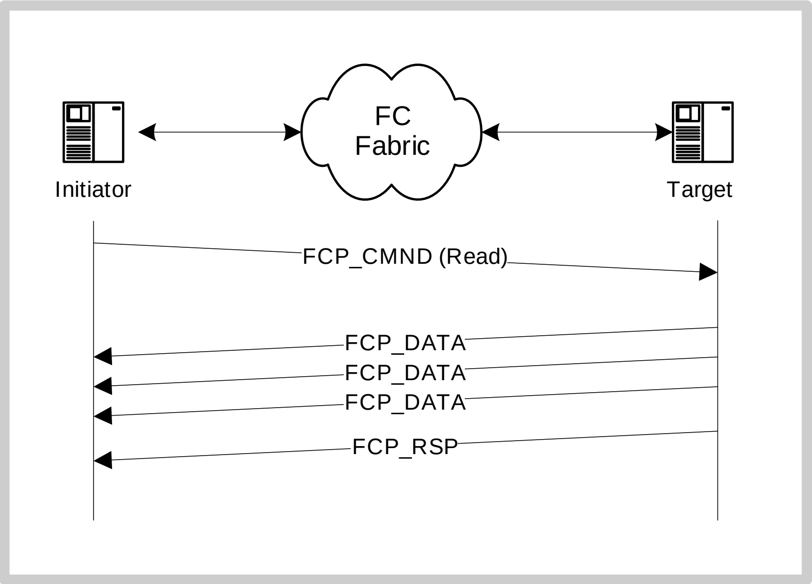
- **Scaling Problem**
  - On ISL State Change,
    - All logical components run
    - Fabric Reconfigurations
      - > Due to dynamic addressing & principal switch election requirements
      - > Halts all traffic through every switch in the Fibre Channel SAN fabric
  - On Device up/down
    - Name Service and State Change Notification Service run
  - As the Fabric grows
    - more and more resources are consumed
    - limits the size of SAN Islands
    - partially solved by deploying multiple SAN Islands
      - > *but, this limits connectivity!*

FC Traffic

- **Most FC Traffic Uses**
  - SCSI-FCP
    - Request-Response Protocol running on top of FC
    - FC provides frame sequencing within transactions
    - Error recovery is at the transaction level
      - > Transactions time out and are retried
      - > Timeouts are preconfigured and not based upon actual RTT
  - FC Class 3
    - Unacknowledged datagram service
    - No individual Frame Retransmission
  - FCP is sensitive to frame loss and errors

FC Traffic

- **READ Command**

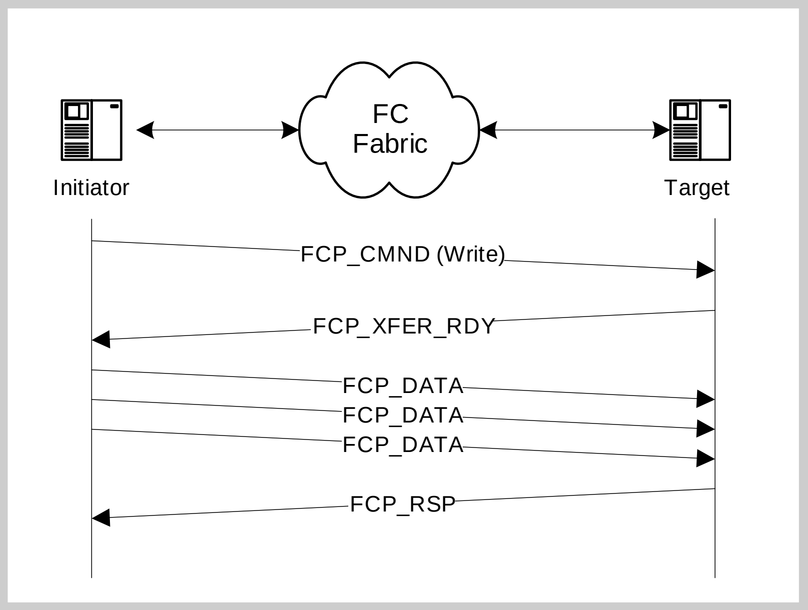


Requires one RTT plus any data transmission delay to complete the command

FCP_CMD (68 bytes)
FCP_RSP (64 bytes)
FCP_DATA (payload size is up to 2048 bytes)

FC Traffic

- **FC Write Command**



Requires two RTTs plus any data transmission delay to complete the command

This fact becomes more important as the bandwidth delay product increases

FCP_CMD (68 bytes)
FCP_XFER_RDY (48 bytes)
FCP_RSP (64 bytes)
FCP_DATA (payload size is up to 2048 bytes)

SAN Deployment Critical Factors

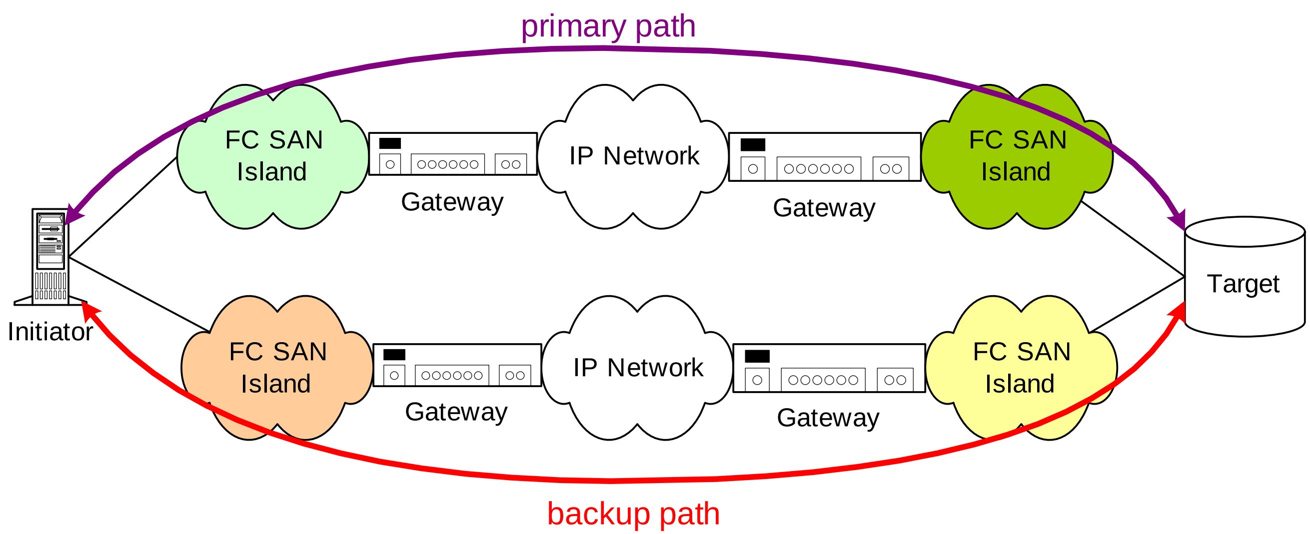
How do the demanding requirements placed on the SAN affect designs?

Critical Factors in SAN Deployments (1)

- **High Availability**
  - Requirements are several ‘9s’ to 99.999% up time
    - Can have requirements for no down time
  - Usually based on a dual-rail configuration
    - All components in the path are replicated
      - > Can be through redundant paths in a single fabric
      - > Can be through completely separate fabrics
      - > Key requirement is no single shared point of failure
    - Individual components are still highly available
  - Director HA Features (and some switches)
    - Fully redundant, hot swappable field replaceable units (FRUs)
    - Hot software download and activation

High Availability

- **A good deployment has parallel paths available**
  - For example, completely parallel networks fill this role
  - Dual ported end devices are also required



- **A poor deployment has a single point of failure**
  - For example if both paths routed through a single gateway in the above picture

Critical Factors in SAN Deployments (2)

- **Robustness and Stability**
  - FC Fabric Build & Initialization Process
    - Heavy Weight mechanisms
      - > Disruptions in one part of the fabric affect the rest
    - Limits the effective FC Fabric size
  - SCSI-FCP
    - Expects and needs low frame loss rates
      - > Some servers, HBAs, and storage devices are very sensitive
      - > Sensitivity extends to frame reordering
    - Error Recovery at the command and transaction level
      - > Time-out and retry mechanisms most widely deployed

Robustness and Stability (cont)

- **Congestion Control**
  - Critical to ensure adequate aggregate bandwidth in designs
  - FC Fabrics use link level credit based flow control
    - good for bursts resulting in short term congestion
    - Active queue management techniques (e.g. RED) not used
  - Frames sitting in the Fabric for too long are discarded
    - typically ½ to 1 second
  - FC moving to more speeds (4G and 10G)
    - creates additional challenges for network and switching architectures
  - Comprehensive congestion management mechanisms

Critical Factors in SAN Deployments (3)

- **Performance**
  - best case latency is a few micro-seconds
    - Applies to most switches and directors
  - latency grows with loading
    - Can be a significant on some switches
      - > At 70% link loading (Miercom Test Lab)
        - Product A : 5.2 us to 6.5 us range
        - Product B: 2.6 us to 2222.6 us range
      - > Longer latencies can have a significant performance impact
        - Not all switches are designed equal!
  - Switching Architecture Issues
    - head-of-line blocking
    - internal throughput bandwidth
    - internal frame rate

Critical Factors in SAN Deployments (4)

- **Distance Extension**
  - Needed for Business Continuance & Disaster Recovery
  - Main Application Activities are
    - File/data Mirroring
    - Data Replication
    - Data Backup
  - Has All of the SAN Island Requirements
    - The manner in which distance interacts with these is critical

Critical Factors in SAN Deployments (5)

- **Scaling the SAN**
  - Most deployments are small, disconnected SAN Islands
    - Still a relatively new technology
    - Practical difficulties with management and stability of large FC Fabrics
    - Business and operational drivers only recently emerged
  - Key Competing Factors in Scaling
    - Want global access between servers and devices
    - Need small stable FC fabrics
  - Compromise by Internetworking the SAN Islands
  - Benefits of Inter-networked Fabrics
    - Resource Sharing (e.g. Backup Tape Library)
    - Dynamic Reassignment of critical resources
    - But, local fabrics can be kept isolated

IP Storage Protocols

FCIP, iFCP, iSCSI, iSNS

07-APR-2004

IP Storage Common Features

- **All IP Storage Protocols leverage the maturity of IP networking**
  - TCP/IP for connectivity
    - Connection Oriented
    - Guaranteed Delivery
    - Congestion Control
    - IP Networks are well understood
    - but… there are issues with the use TCP
  - Inherently allows for usage of IPSec
    - Authentication
    - Integrity
    - Privacy

iFCP & FCIP Common Encapsulation

- **FC frames from local FC devices are encapsulated with a protocol specific header**
  - Conforms to the common encapsulation header format from the IETF FC Frame Encapsulation draft standard
  - The encapsulated frames are then sent across the IP network on the protocol appropriate TCP connection
  - The frames are de-encapsulated at the remote gateway and sent to the remote FC device through the remote FC Fabric

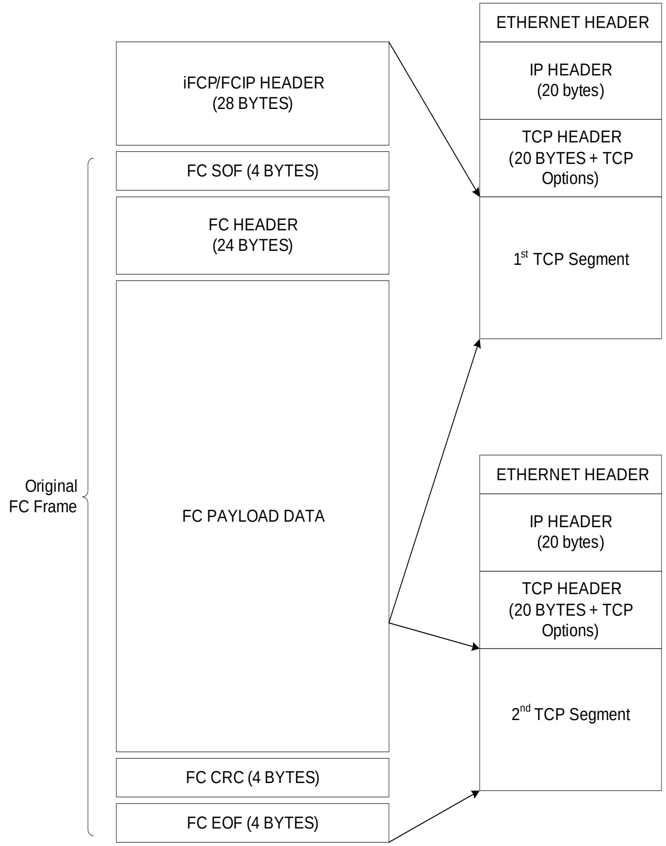
Word (bits 31 … 0)				
0	PROTOCOL	VERSION	-PROTOCOL	-VERSION
1	copy of word 0			
2	pFlags	reserved	~pFlags	~reserved
3	ENCAP FLAGS	FRAME LENGTH	-ENCAP FLAGS	-FRAME LENGTH
4	TIME STAMP (INTEGER)			
5	TIME STAMP (FRACTION)			
6	iFCP HEADER CRC			

FCIP Header

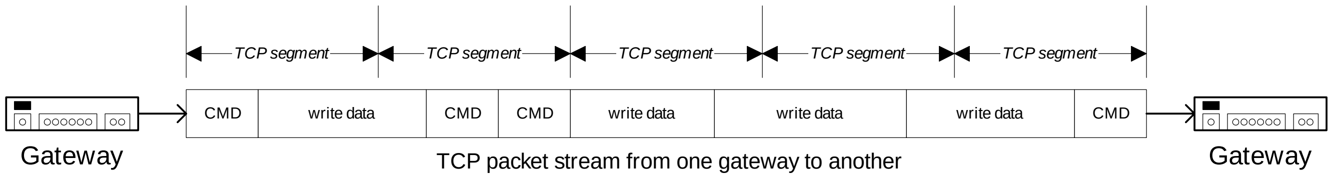
Word (bits 31 … 0)				
0	PROTOCOL	VERSION	-PROTOCOL	-VERSION
1	RESERVED (MUST BE 0)			
2	LS_COMMAND	FLAGS	SOF	EOF
3	ENCAP FLAGS	FRAME LENGTH	-ENCAP FLAGS	-FRAME LENGTH
4	TIME STAMP (INTEGER)			
5	TIME STAMP (FRACTION)			
6	iFCP HEADER CRC			

iFCP Header

iFCP & FCIP Common Encapsulation



FC Frame Encapsulation
and
TCP Segmentation



FCIP

- **IETF Standards Track Specification as part of the IP Storage Working Group**
  - Connects local FC Regions together over an IP Network into a single unified FC Fabric
  - FC frames are transparently encapsulated and tunneled across the IP Network to the other fabric
    - Sends all FC ISL traffic across the tunnel
  - FCIP appears to be a long latency E-Port to the FC Fabric
    - The connection is implemented through FCIP endpoints providing either B-Port or E-Port interfaces into the local FC fabric and using the FCIP protocol itself to connect across the IP Network (FC-BB-2 specification)

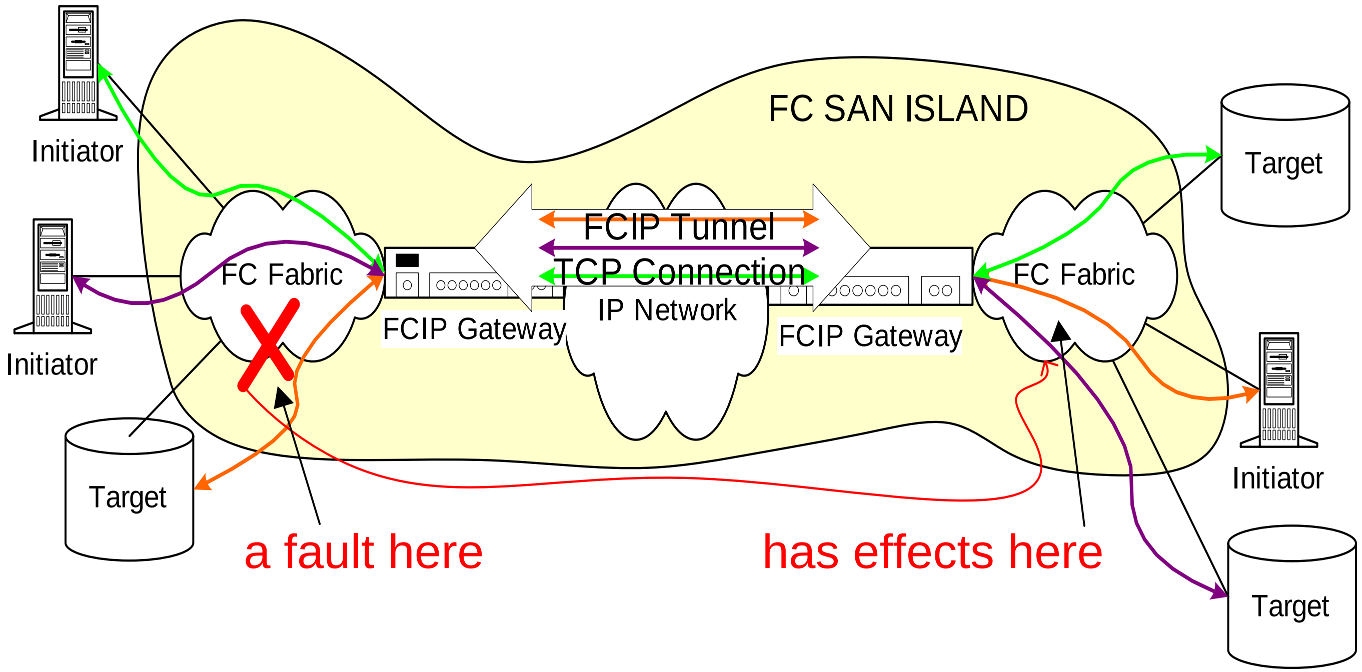
FCIP Advantages and Disadvantages

– Biggest advantage is transparency to Fabric
  - Existing Fabric Tools and Services are used
  - FSPF, Fabric Initialization, Name Service, State Change Notification Service all run transparently

– Must understand
  - Design and management of congestion and overload conditions
  - Tunnel queuing design
    - > to avoid head-of-line blocking of fabric initialization protocol frames when congestion does occur
    - > to avoid congestion collapse due to command timeout
  - These concerns are less important in a small fabric

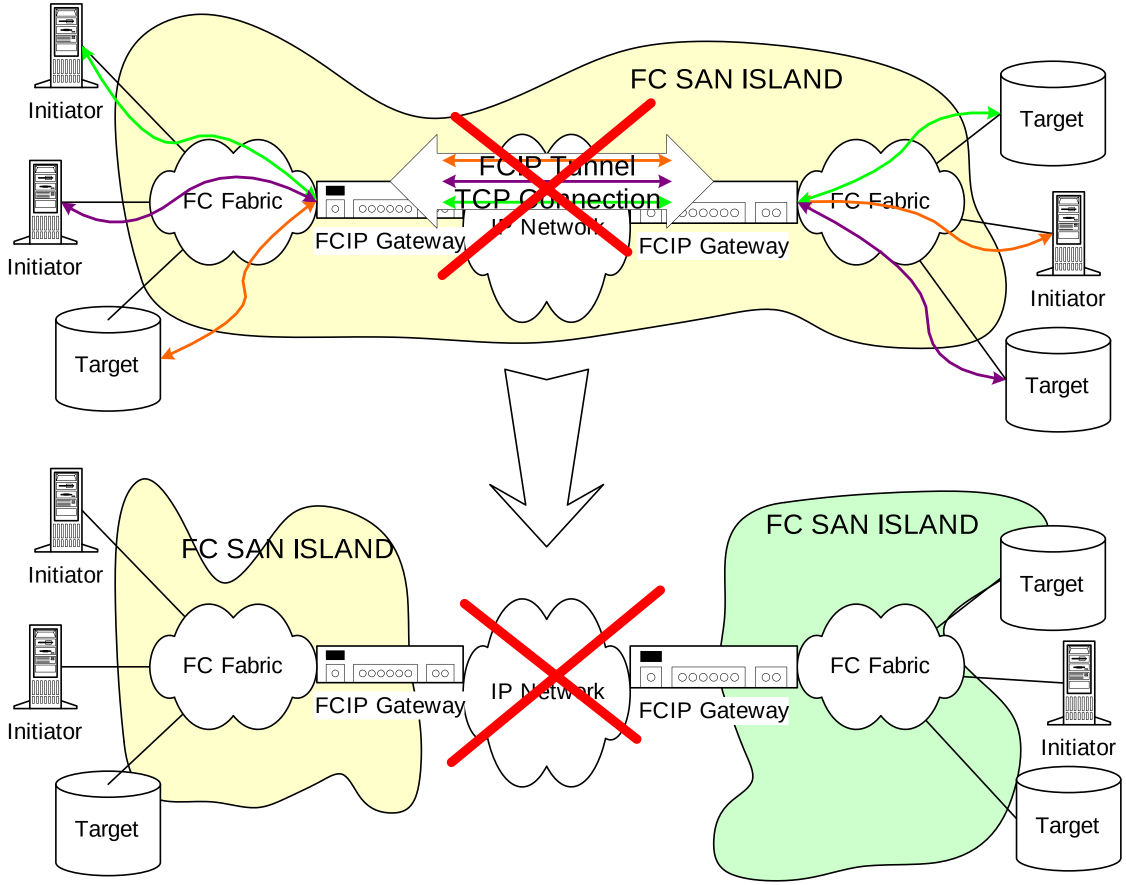
FCIP Connection Example

- **All FC sessions are tunneled across the FCIP TCP Connection**
- **The two local FC fabrics act as a unified FC SAN Island**
- **This provides for transparent WAN interconnection at the cost of fault propagation**



FCIP WAN Disruption

- **Original Fabric Segments into two pieces**
- **Each piece reconfigures into a new SAN Island**



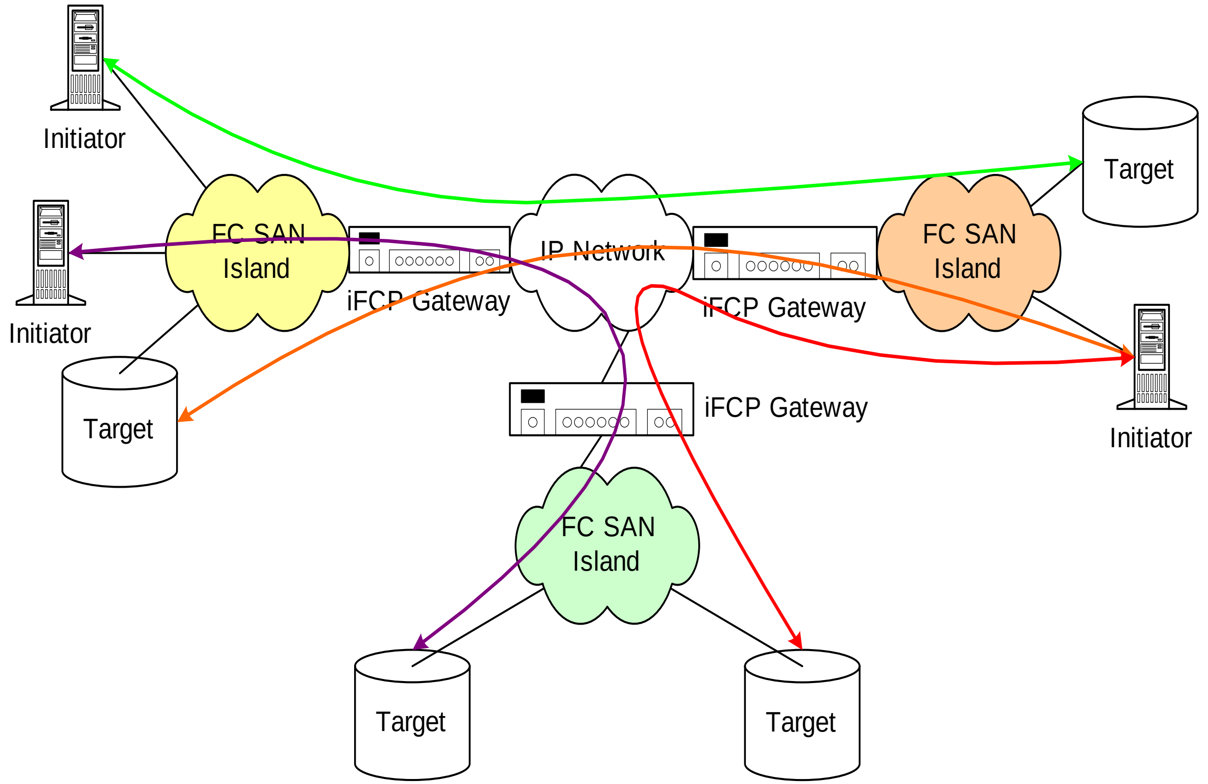
iFCP

- **IETF Standards Track Specification as part of the IP Storage Working Group**

- **iFCP is an FC region interconnect protocol that provides:**
  - FC device communication across an arbitrary IP network using TCP/IP
  - Fault isolation between the autonomous FC regions
  - Each device pair has a separate TCP connection
  - Each FC region maintains its own Name Server
  - iFCP operates between gateways in each FC region.

iFCP Session Example

- **Each of the sessions runs on its own TCP connection**

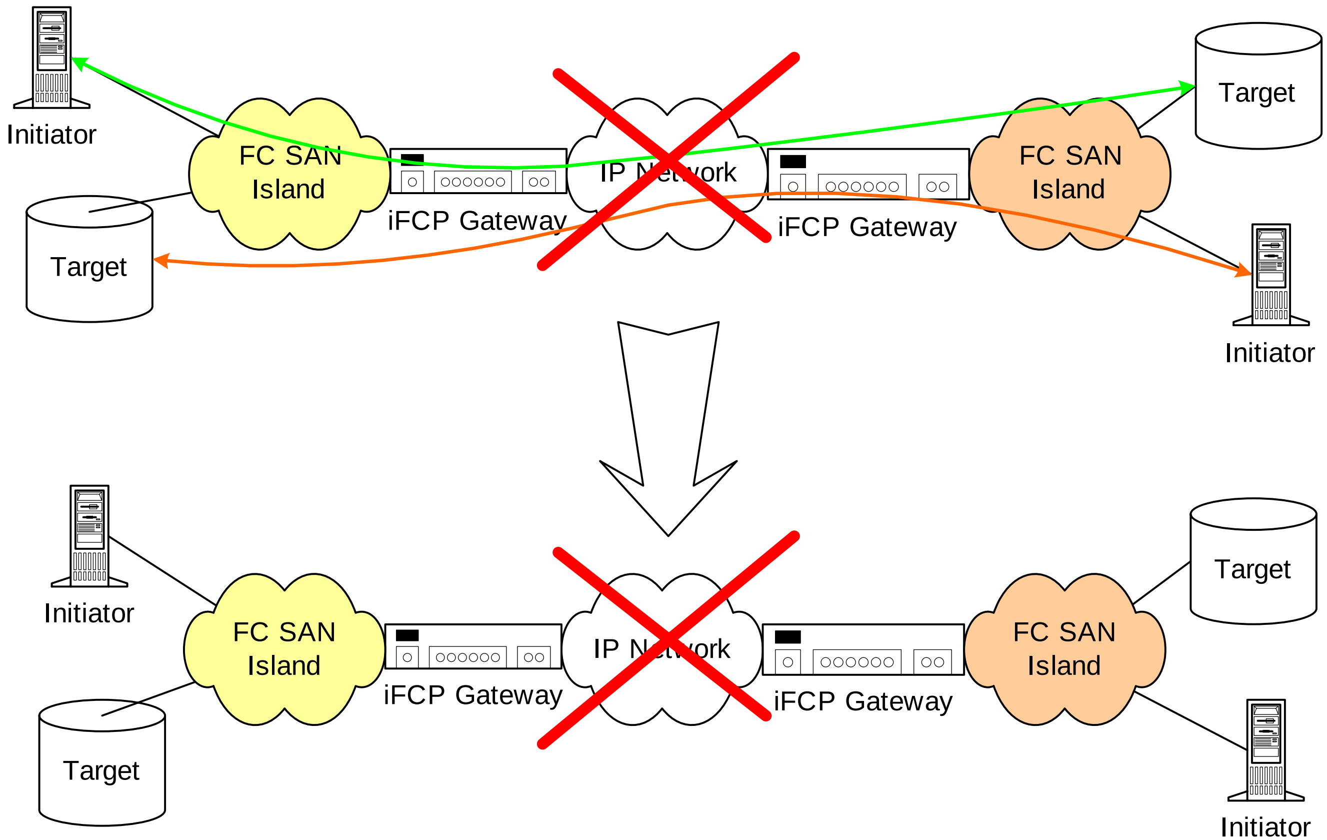


iFCP

- **iFCP Gateways:**
  - Are transparent to FC frame traffic
    - Manage each session’s state
    - Acts as a local proxy for the remote device
  - Provide fabric services for remote devices into the local fabric
  - Intercept and filter local fabric services traffic to the remote FC region
  - Addressing Modes
    - Translate local FC addresses to/from IP addresses and TCP port numbers
    - Use shared addresses for the connected FC SAN Islands (faults are still isolated)
  - Set the policy for FC device export/import

iFCP WAN Disruption

- **Remote Devices are deregistered generating SCNs to each local fabric**



iSCSI

- **iSCSI is a Native IP Storage Block Level Protocol**
  - Designed for use directly by end devices
  - Commands still conform to T10 SCSI specifications
  - Uses TCP as its Network transport
  - Uses IPSec for Security
  - Not an FC encapsulation protocol

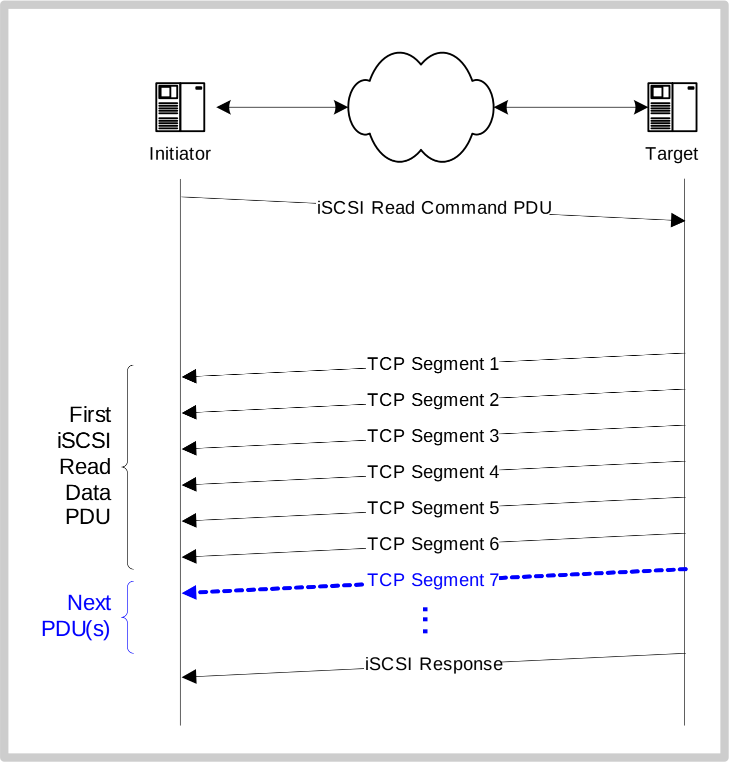
- **iSCSI Protocol Features**
  - Multiple TCP connections for a single session
  - CRC digests for header and data
  - TCP connection failure recovery options
  - Out of Order data placement
  - Supported Options Negotiated on Session Creation

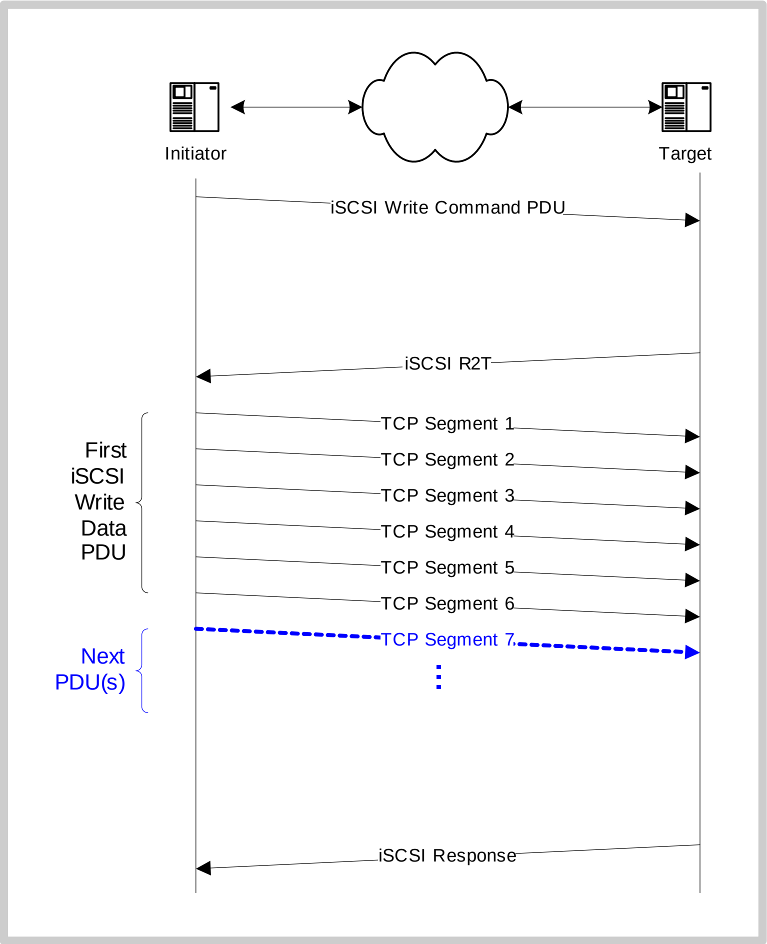
iSCSI Options

A wide variety of negotiable Options exist including:

- Security Parameters
- Large PDU
- Store & Forward
- Header & Data Digests
- Target Read Padding
- Target Write Padding
- NOP-In PDU transmit
- Immediate Data Allowed
- Authentication Support
- Initial R2T
- Multiple R2T
- Asynchronous Messages
- Multiple TCP Connections per Login Session
- SNACK
- Error Recovery Level
- and more…

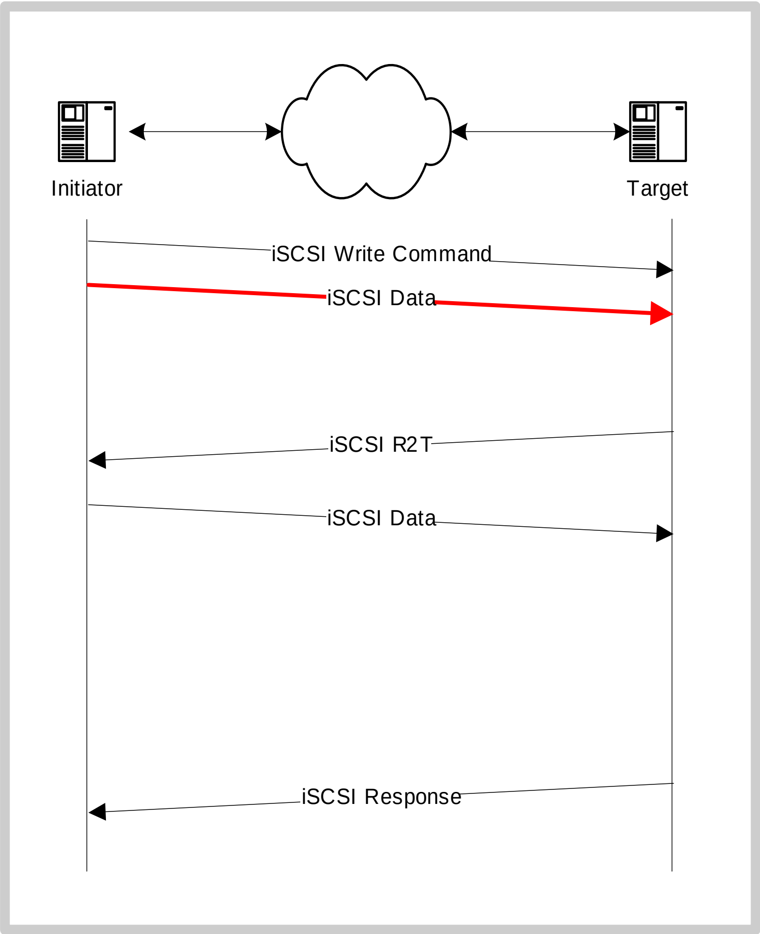
- **Command and Data Flow**
  – Almost the same as FCP command and data flow





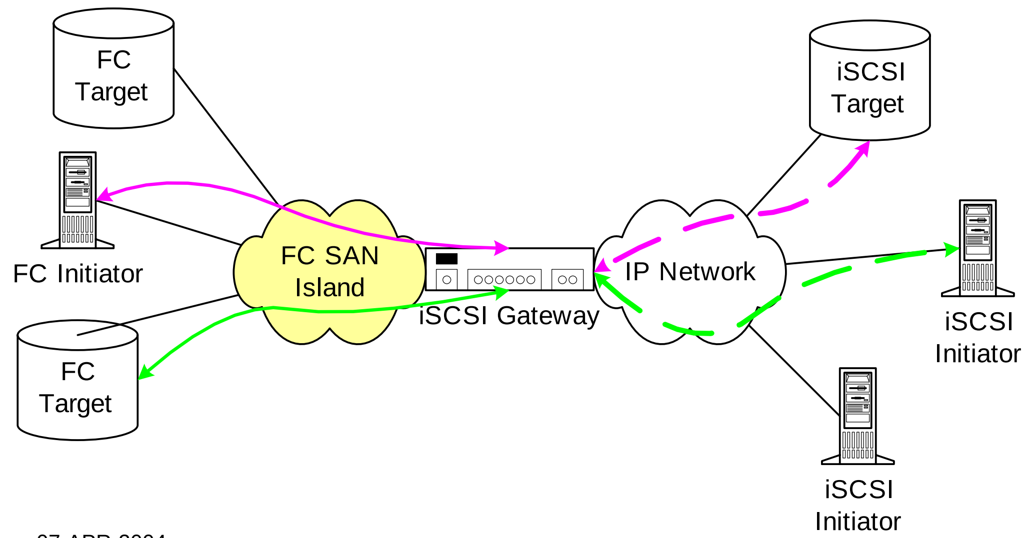
iSCSI

- **iSCSI also provides for unsolicited write data**



iSCSI

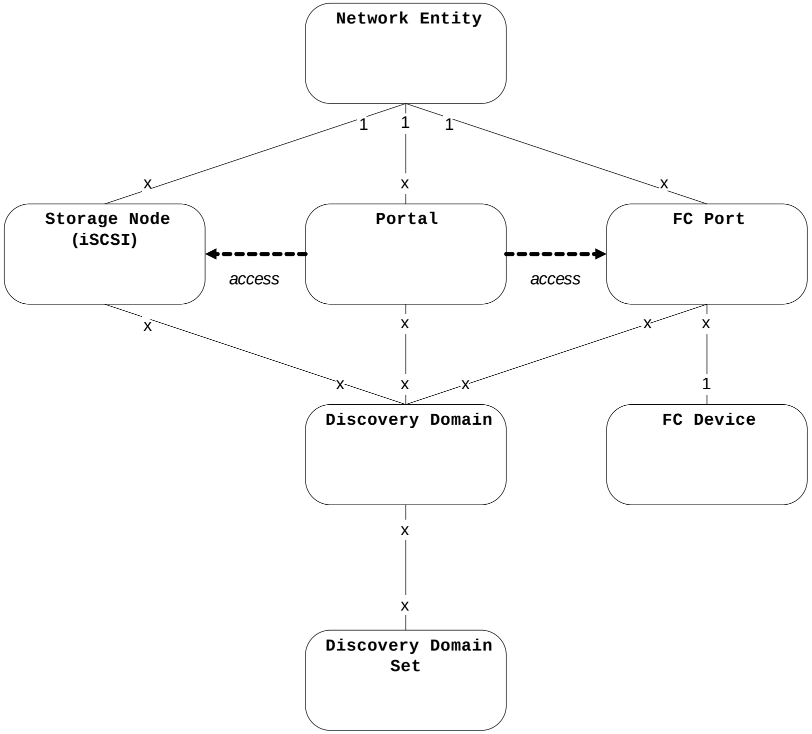
- **FC – iSCSI Gateway**
  - An FC session is established between the gateway and the FC device
  - An iSCSI session is established between the gateway and the iSCSI Device
  - The gateway acts as a proxy of the appropriate protocol into each network
  - performs
    - session level protocol translations
    - command and payload forwarding
    - Device import, export, and mapping
      - > (protocol appropriate name service registrations)
    - login and session creation
  - must maintain
    - Transparency
    - Data Integrity
    - Throughput
    - I/O Rate
    - Sufficient Sessions
      - > Storage Consolidation



iSNS

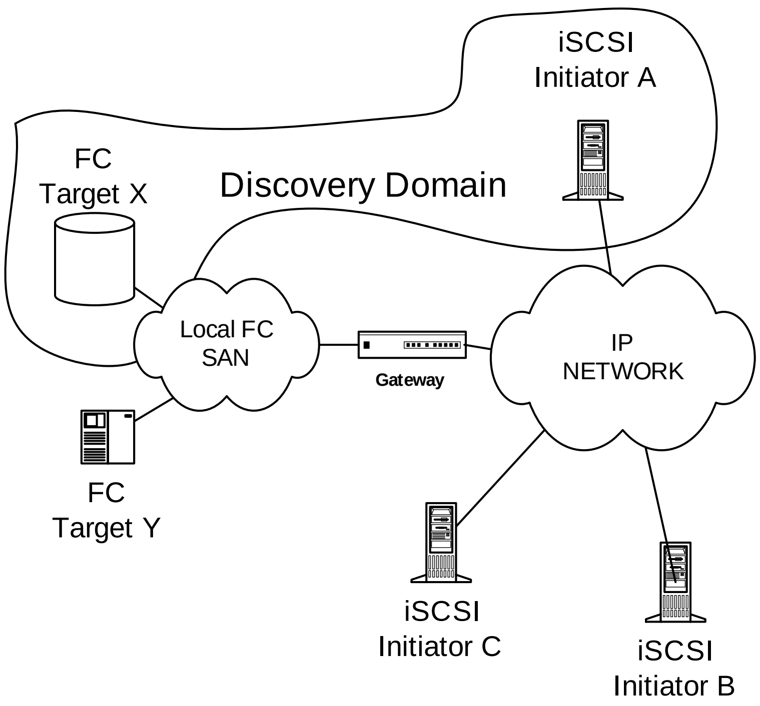
- internet Storage Name Service
- IETF standards track in final review to be ratified soon after iSCSI
- Centralized database for IP Storage device discovery
  - Supports both iSCSI and iFCP
- Filtered asynchronous notification of state changes
- Scoped query support
- Controlled directly through the iSNS Protocol or indirectly through SNMP
- TCP/IP used for client connections
  - UDP support is optional
- Extensible object and attribute scheme

iSNS Object Model



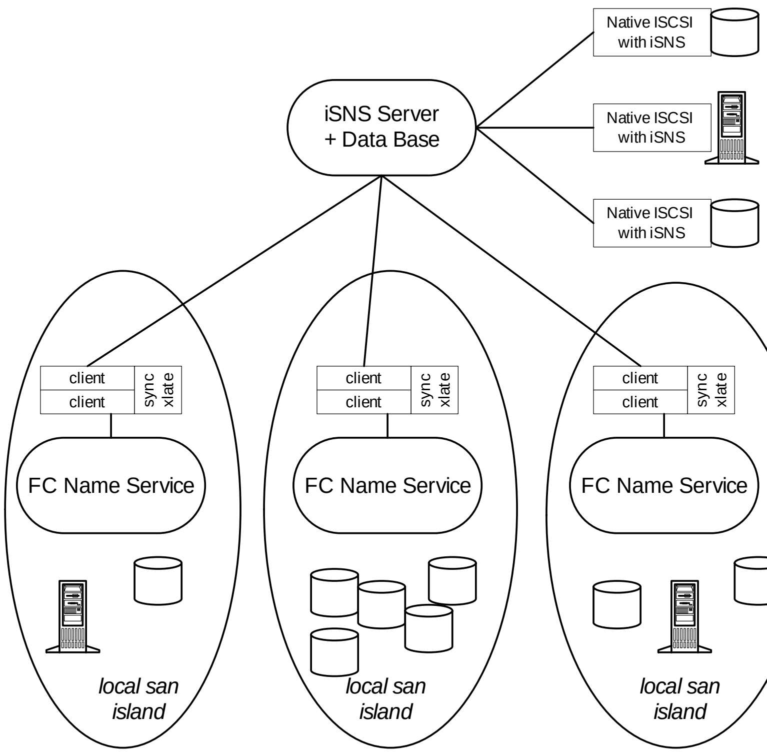
- iSNS Objects match the iSCSI object model
- For iFCP, the objects are their iSCSI equivalents
- The Network Entity implies a physical grouping of nodes, portals, and FC ports
- Within an Entity any portal can be used to access any Node or FC Port
- Discovery Domains (DD) and DD Sets are the iSNS analog to Zones and Zone Sets and allow logical groupings for discovery purposes

iSNS - Discovery Domain Example



- iSNS allows end devices to determine who they may be able to communicate with through the use of Discovery Domains (DD).
  - A DD is created and populated with devices by a control node
  - A device can be in multiple DD's at the same time
  - With asynchronous State Change Notifications (SCN's), devices can be informed of any change in a DD they are in.
- Initiator A can find out about Target X by sending a query to the iSNS Server. Since X is in the same DD as Initiator A, Initiator A will be informed about the registered device information for Target X. Initiator A can use this information to login to X.
- The iSNS Server will not inform Initiator A about Target Y, Initiator B, or Initiator C.
- If Target Y were subsequently put into the DD, Initiator A would receive an SCN event informing it that Target Y had been added.

iSNS in a heterogenous environment



- One iSNS Server per network/SAN.
- One current SNS instance per local SAN island (1 switch to a few switches).
- A synchronization and translation task in each local SAN takes care of registering "external devices" in the local SNS and local devices in the central iSNS.
- Registrations use the appropriate addressing and gateway view for the database being used.
- To the iSNS server and data base, each local SAN island looks like a multi-portal and multi-storage node iSCSI device with native iSNS clients and devices.
- To the SNS, external iSCSI devices look like locally attached mFCP devices.

Case Studies and Issues

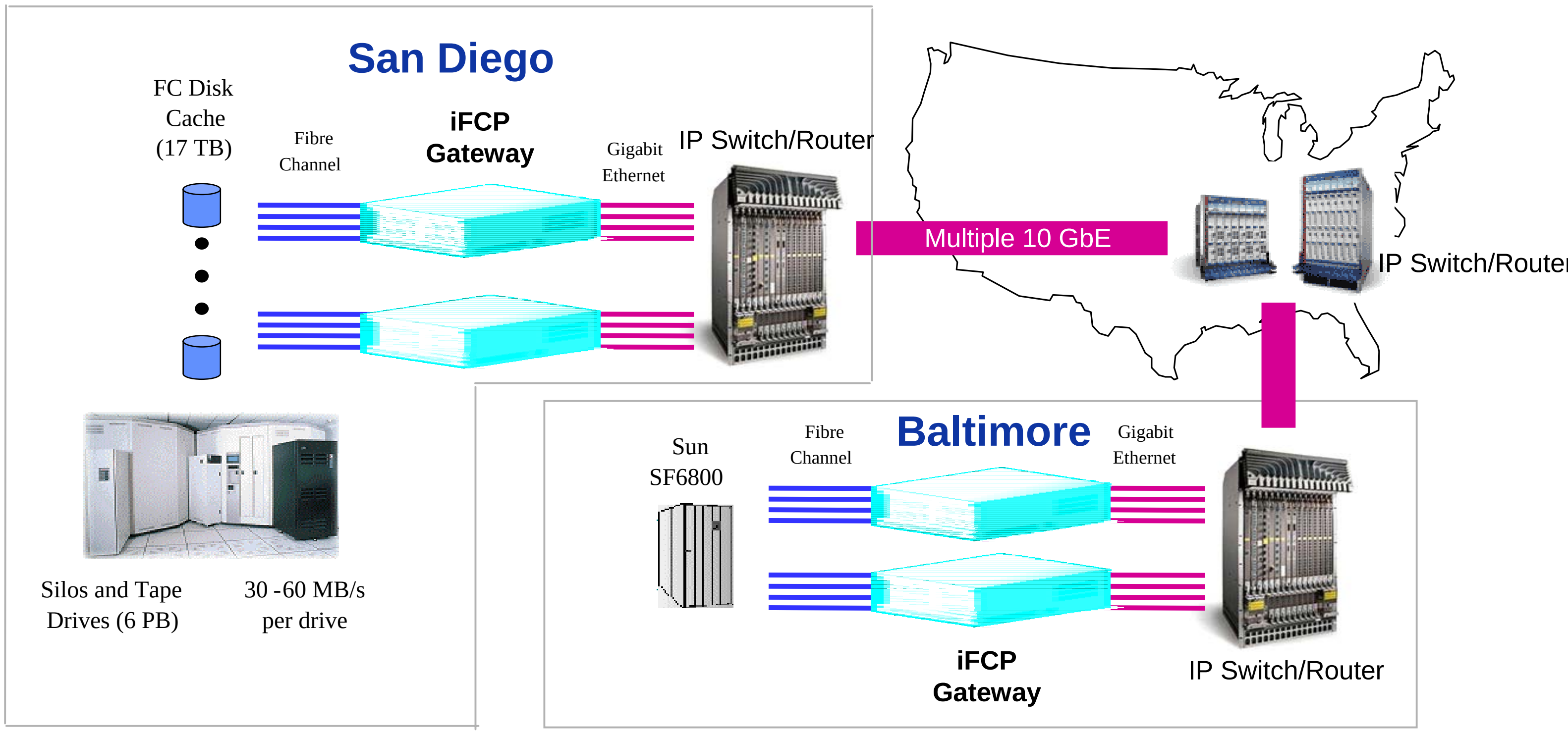
What are the difficulties in converging the SAN?
What are some solutions to those difficulties?

TCP Performance

- **Ongoing TCP Research in a variety of areas**
  - Operation in High Speed, Long Latency Networks
    - SACK (Selective Acknowledgement)
    - ECN (Explicit Congestion Notification)
    - Eifel Detection Algorithm
    - HSTCP (High Speed TCP)
- **TCP for SAN**
  - use of large receive windows for distance connections
  - multi-connection IP Storage Sessions
  - buffer-MTU-MSS tuning
  - TCP offload engines
    - plus other iSCSI driver specific enhancements
- **Additional efforts exist for Ethernet and FC directly over SONET-based transports**

SDSC: GRID Computing IP SAN

San Diego Supercomputer Center



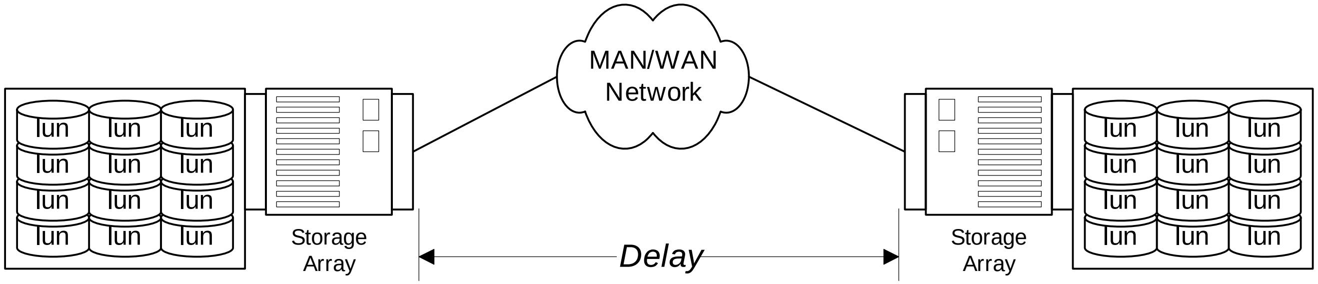
Sustained Aggregate Performance: ~717 MBps

iFCP from San Diego to Baltimore Fall 2002

San Diego Supercomputer Center

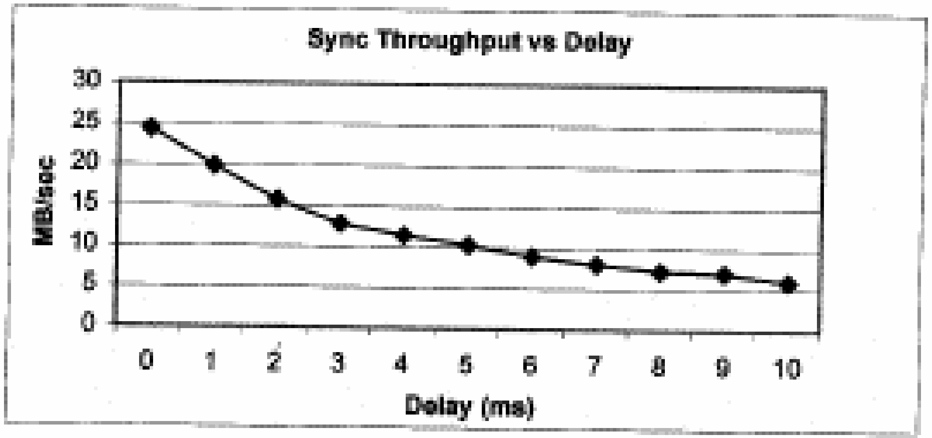
- **Test was part of the 2002 Supercomputing Conference**
- **Total distance was 2600 miles**
- **RTT was 70 to 90 ms**
- **FC traffic using iFCP gateways to interconnect SANs**
  - interconnect was a 10 Gb/s
- **Aggregate Sustained Performance of 717 MB/s**
  - Read throughput was 740 MB/s
  - 8 x 1G IP Gateways used in parallel
    - about 90MB/s sustained throughput per gateway

Remote Mirroring



- **The remote copy synchronizes the data copied between each LUN pair**
  - Ensures data coherency within the mirror group
  - Each LUN pair is allowed 1 outstanding command
  - Command must complete before data is committed

Remote Mirroring

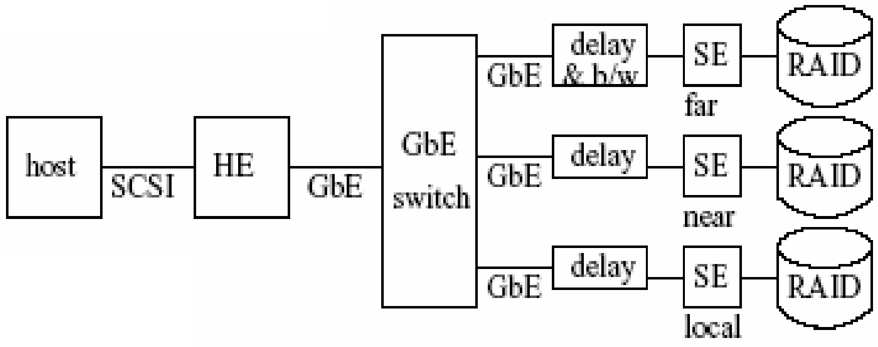


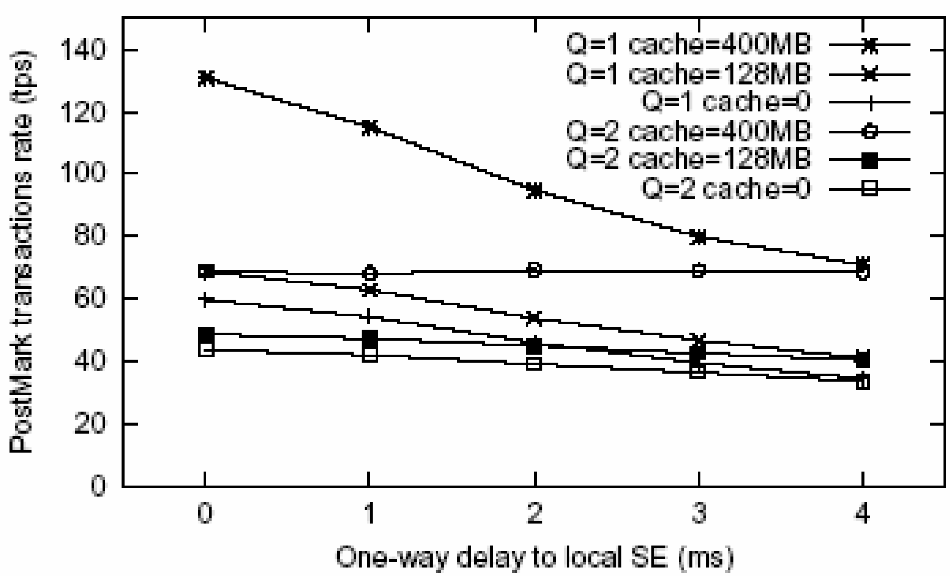
Effect of Network Delay on Synchronization Throughput

- **The throughput in this configuration fell off rapidly with latency**
- **Parameters affecting throughput**
  - file and write command transfer size
  - Command type
    - (write commands take 2*RTT to complete)
  - number of outstanding commands
    - The bandwidth delay product must be filled for max throughput
  - In a low bandwidth environment compression is critical

Delay and Cache Effect on I/O Workload

- **StarFish:**
- **Test the delay and cache sensitivity of a server**
  - PostMark used as test program
  - Large E-Mail server attached to Storage Elements (SE)





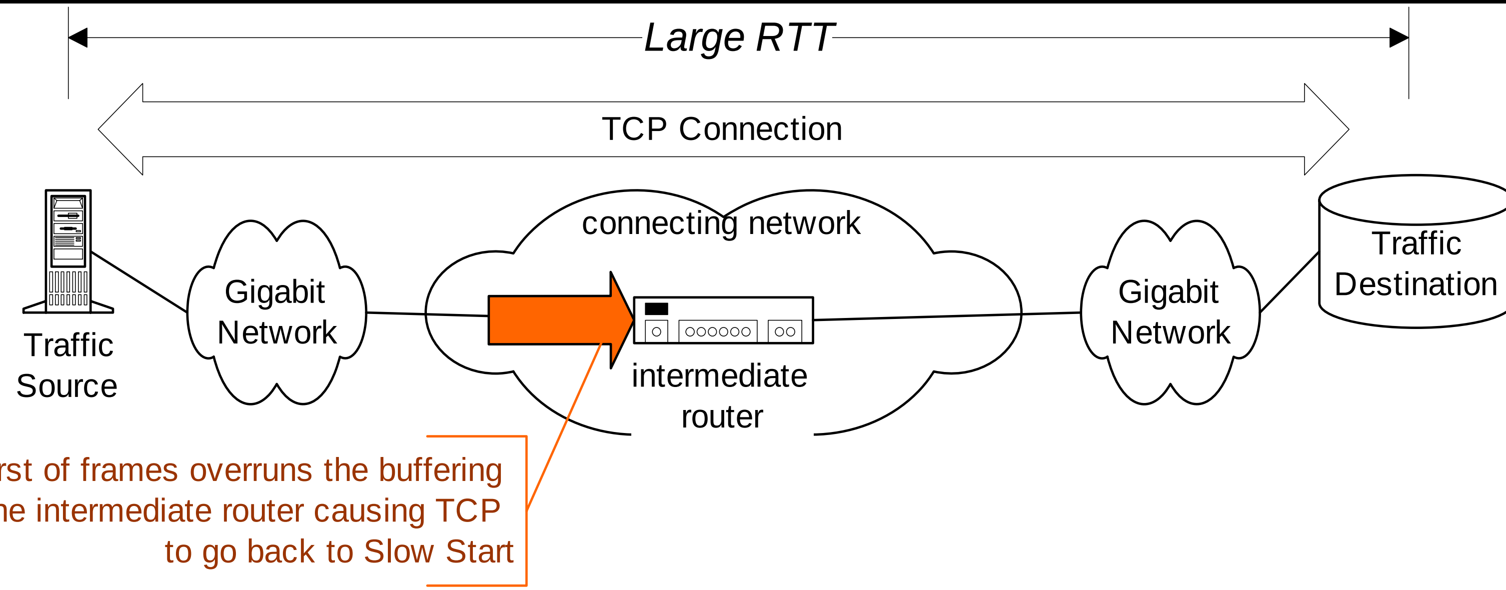
Delay and Cache Effect on I/O Workload

- **<u>Observations:</u>**
  - As delay increases the transaction rate supported by the server decreases.
  - As the server cache increases the transaction rate increases.
  - Writes are not helped by the cache in their architecture so the distance affects writes more.
    - The Q1 curves are to the local SE (with varying latency)
    - The Q2 curves also involve a second ‘near’ SE with little latency. This improves the write performance.
- **Observations are not surprising but several places exist for performance improvements**
- **Application performance sensitivity wrt delay and error recovery is a complex area that needs further work and understanding**

Long Fast Network Experiment – U. of Tokyo

- **iperf run between Maryland and Tokyo**
- **Produced the counterintuitive result that Fast Ethernet is "faster" than Gigabit Ethernet in some cases**
- **Conclusion is that transmission rate limiting is important in sustained TCP throughput across lower speed non-flow controlled networks**

Long Fast Network Experiment – U. of Tokyo



- **TCP factors contributing to slower than expected performance**
  - The rate at which the congestion window ramps is determined by the RTT
    - This acts a limit on the transmission rate since only one congestion window worth of data can be in flight per RTT
  - When missing frames can't be recovered with fast retransmission, the connection will eventually time those frames out and drop back into slow start
    - In slow start the congestion window is reset and must grow again
    - This timeout value is at least 1 second and can be longer if the RTT is large enough

Fast Write

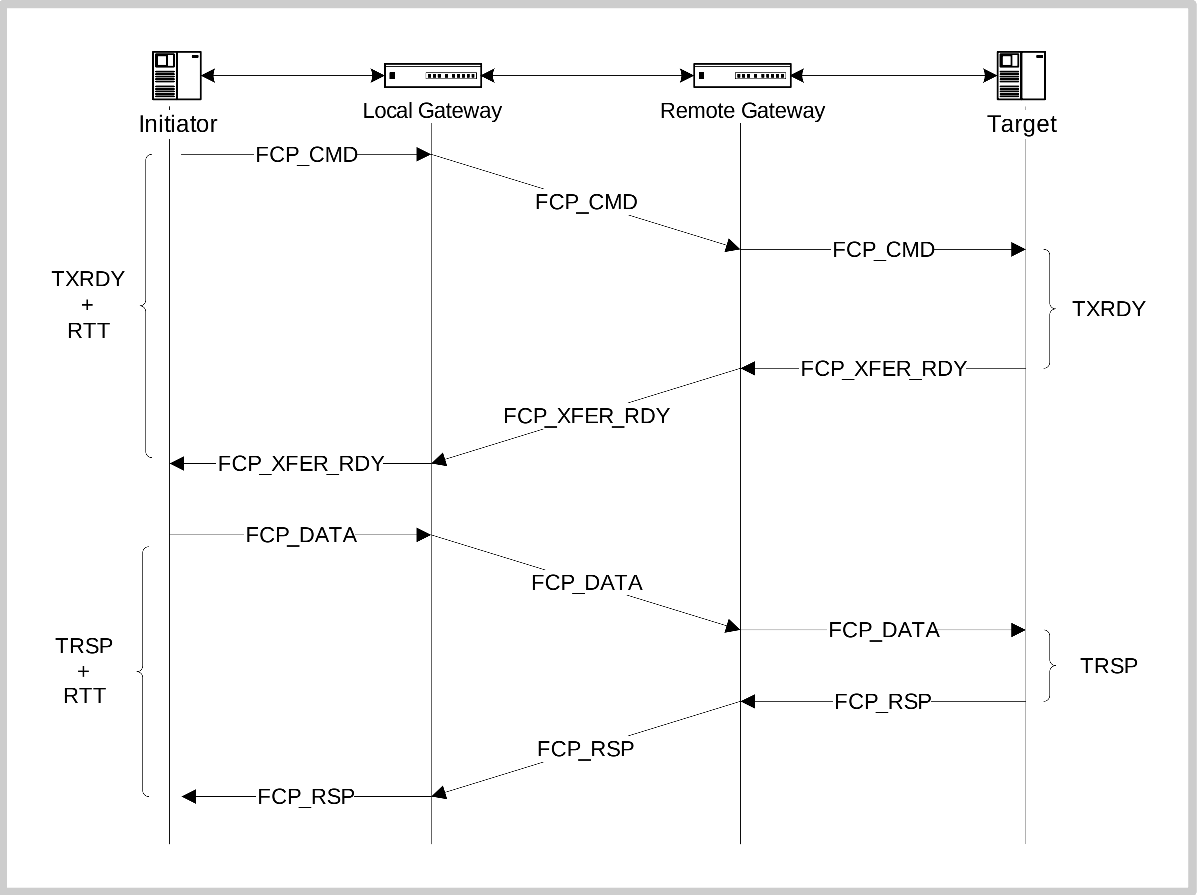
- **A method to improve write performance in a long delay network**

- **Reduces the number of RTTs needed to complete a write command from 2 to 1**
  - In other words, this makes the effective network latency look like it has been cut in half for write commands

- **Not a command completion spoofing algorithm**
  - The FCP_RSP still flows from the target back to the initiator

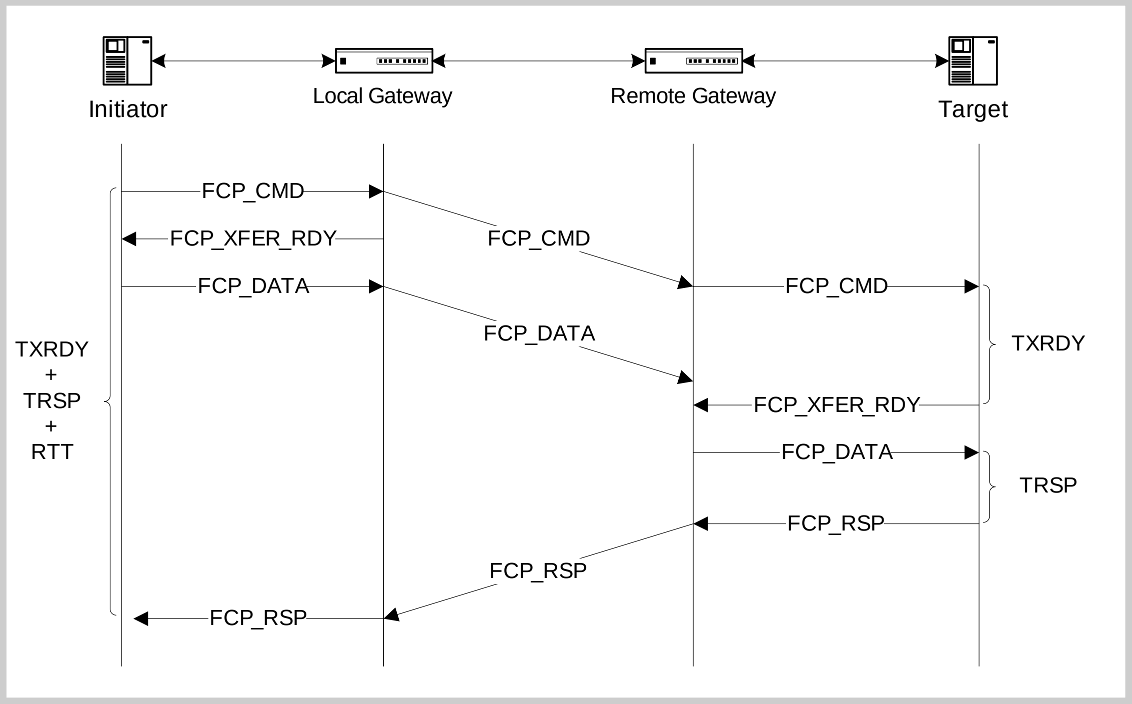
Fast Write

- **Write Command Latency**
  - Increases by one additional RTT (Network Round Trip Time) per XFER_RDY by the target
  - At least 2*RTT (a read command would only be one RTT)



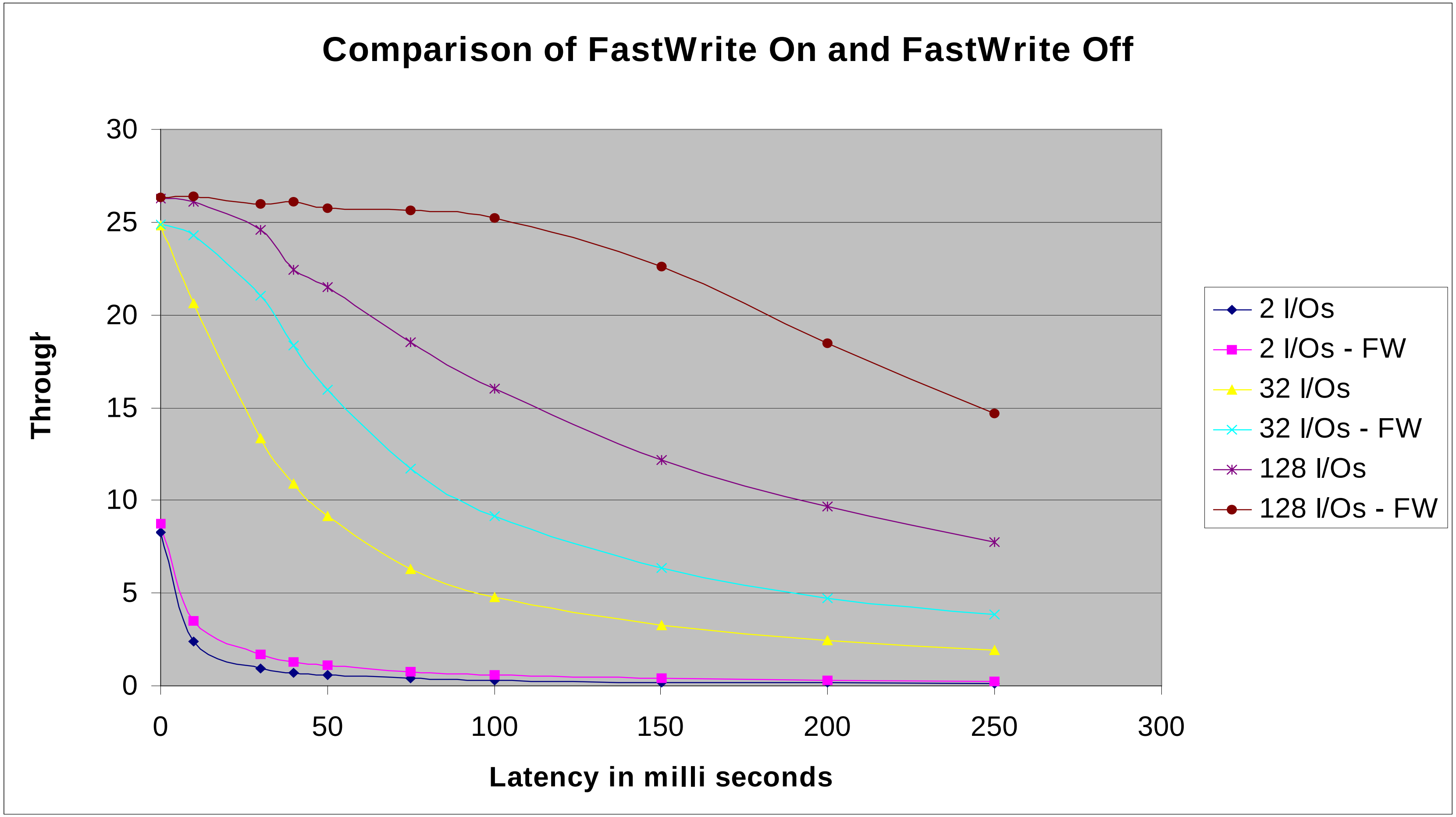
Fast Write

- **Fast Write Reduces Write Command Latency to 1 RTT**
  – Timing is now similar to a read command



- ***Reduces the Effective Link Latency by 50% For Write Commands***
  – Better Performance Over High Latency Link
  – 50% reduction assumes a Single XFER_RDY Issued by Target
  – Effective Latency Reduced Even More if Target Issues Multiple XFER_RDY for a Command

Fast Write



Comparison of FastWrite On and FastWrite Off
Througr
30
25
20
15
10
5
0
0
50
100
150
200
250
300
Latency in milli seconds
2 I/Os
2 I/Os - FW
32 I/Os
32 I/Os - FW
128 I/Os
128 I/Os - FW

Conclusions and Questions

Conclusion

- **IP and FC SAN integration can work well**
  - Even in the face of the stringent SAN requirements
  - Proven in real world deployments
  - More integration is inevitable

- **Challenges still exist for many converged and IP Storage deployment scenarios**
  - This is especially true of long distance and high speed connections
  - Lots of different work exists or is underway to solve the problems
  - Can’t ignore FC SAN requirements

- **Despite the challenges integrated and converged deployments are very promising with a large amount of practical value**

Questions and Discussion

Questions?